De novo assembled single-cell transcriptomes from aquatic phytoflagellates reveal a metabolically distinct cell population

Javier Florenza[1,*], Aditya Jeevannavar[2,*], Anna-Maria Divne[3], Manu Tamminen[2,*], Stefan Bertilsson[1,4,*]

[1] Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden, [2] Department of Biology, University of Turku, Turku, Finland, [3] Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden, [4] Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden.

[*]These authors contributed equally to this work.

Abstract

Single-cell transcriptomics have rapidly become a standard tool for decoding cell identity, fate, and interactions in mammalian model organisms. Adopting these techniques to uncover functional dynamics in aquatic single-celled organisms holds huge potential, but evidence of their applicability to non-model, poorly understood microeukaryotes remains limited. In the present study, live *Ochromonas triangulata* cells from fast and slow growth phases were FACS-sorted based on food vacuole staining and chlorophyll fluorescence, and single-cell transcriptomic libraries were prepared following the Smart-seq2 protocol. In total, 744 transcriptomes were Illumina sequenced. Lacking a reference genome, transcriptomes were assembled *de novo* using Trinity and resulting transcripts were annotated by BLAST using the Swiss-Prot database. Following read mapping, differential gene expression was evaluated using DESeq2 and metabolic maps were generated based on pathways from the KEGG Orthology database. Clustering the read counts revealed the identity of the two expected transcriptional states corresponding to each growth phase as well as a third distinct cluster of cells present in both growth phases. This cryptic group showed extensive downregulation of genes in pathways associated with ribosome-functioning, CO_2 fixation and core carbohydrate catabolism such as glycolysis, β oxidation of fatty acids, and tricarboxylic acid cycle. Nevertheless, the biological underpinnings of this cluster, which would have remained unnoticed in an integrated approach, could not be clarified. Additionally, the possibility of using carry-over rRNA reads for taxonomic annotation was tested, which revealed specific individual-level bacterial communities associated with members within each cluster, as well as verified identification of 88% of the cells as *O. triangulata*. In conclusion, we demonstrate the power of single cell transcriptomics for metabolic mapping of microeukaryotes for which genomic reference resources might be limited, and thereby highlight its potential as a tool to gain access to microeukaryote dynamics and diversity in natural communities.

Keywords: single-cell transcriptomics, Smart-seq2, Ochromonas triangulata, SSU rRNA, cryptic diversity

Introduction

Microbial eukaryotes comprise the vast majority of known eukaryotic lifeforms (Simpson et al. 2017). They showcase a plethora of lifestyles, from free living to strictly parasitic (Walker et al. 2011), photosynthetic, phagotrophic or both (Worden et al. 2015, Mitra et al. 2016). They exhibit a varied catalogue of gene expression quirks, such as mRNA fragmentation, *trans*-splicing or translational slippage (Smith and Keeling 2016). They are ubiquitous across various aquatic and terrestrial environments (Geisen et al. 2015, 2016, Pernice et al. 2016), ranging from freshwater to extreme habitats (Iniesto et al. 2022, Jamy et al. 2022), tropical to polar (Marshall and Laybourn-Parry 2002, Mahé et al. 2017), and are furthermore subject to very strong spatial (Forster et al. 2016) and seasonal constraints (Moreira and López-García 2019, Vass et al. 2020). The unifying feature that groups them together, beyond their eukaryotic character, is that they are single celled and, therefore, microscopic. Their small size, coupled to the difficulty of culturing the overwhelming majority of known single-celled eukaryotic organisms (Del Campo et al. 2014), has technically limited the study of their taxonomy, physiology and ecology to those species that could be readily cultured or visually recognized under the microscope.

The advent and subsequent development of techniques based on nucleic acid sequencing has circumvented these limitations, greatly expanding our understanding of microeukaryote biology (Goffeau et al. 1996, Venter et al. 2004). During the last two decades, several *en masse* approaches based on molecular characterization (i.e. "omics" methods) have developed into pivotal tools to study the functional potential of microbial species and the composition of microbial communities. Whenever technically feasible, genomic and transcriptomic exploration have become a cornerstone in microeukaryote research (Del Campo et al. 2014, Keeling et al. 2014). When cultured specimens are not available, or the focus is on microeukaryotes in the wild, metagenomics, in either gene barcoding (Massana et al. 2014) or whole-genome fashion (Lepere et al. 2011, Mangot et al. 2017), can be deployed to investigate taxonomic and functional diversity, respectively. Additionally, metatranscriptomics has yielded valuable insights about the physiology of densely populated phytoplankton communities (Marchetti et al. 2012, Cooper et al. 2014).

All these approaches, however, lack single-cell resolution and therefore cannot capture heterogeneity at organismal level. Single-cell methodologies have emerged to overcome that. These were first successfully applied to recount clonal expansions in single nuclei of human tumours (Navin et al. 2011), describe somatic diversity in the genomes of human neurons (McConnell et al. 2013) and detect gene expression heterogeneity in human leukocytes (Mahata et al. 2014), and were quickly adapted to target functional and taxonomic heterogeneity in environmental samples, both for prokaryotes (Rinke et al. 2013) and eukaryotes (Yoon et al. 2011). The advantages of targeting

individual cells are multiple: assays require little starting material, are independent of culture availability, are not limited by high microbial community diversity and provide resolution at organismal level. In the case of single cell mRNA sequencing (scRNA-seq), single cell resolution allows accessing transcriptomic information that would otherwise be too scarce to be recovered from bulk extraction (for example, when the abundance of target organisms is low), and can resolve coexisting, conspecific cell types with disparate gene expression profiles which otherwise would be averaged in an integrated sample.

Despite this potential, studies targeting gene expression in microeukaryotes at single cell levels are scarce, and mostly committed to either yeast model organisms (Nadal-Ribelles et al. 2019, Saint et al. 2019) or pathogenic apicomplexans, such as *Plasmodium* (Poran et al. 2017, Reid et al. 2018, Howick et al. 2019) and *Toxoplasma* (Xue et al. 2020). To date, only two publications exist that offer insight into microeukaryote gene expression beyond either of these lineages: one exploring the feasibility of expression-targeted scRNA-seq in two small flagellates, the haptophyte *Prymnesium parvum* and the dinoflagellate *Karlodinium veneficum* (Liu et al. 2017) and another uncovering successive stages of infection in the haptophyte *Emiliania huxleyi* by its specific giant virus (Ku et al. 2020).

This paucity of studies is partly a consequence of methodological constraints and challenges associated with single cell transcriptomic techniques, such as the need for rapid cell isolation upon sampling, low throughput of isolation strategies compatible with large cells where manual cell isolation is the only feasible option, the low number of mRNA molecules per individual cells or the difficulty of lysing walled or shielded cell types (Liu et al. 2017, Ku and Sebé-Pedrós 2019). Despite these limitations, the *Plasmodium* and *Emiliania* examples showcase scRNA-seq as a feasible approach to metabolically map distinct cell stages within complex microbial populations. However, both examples rely on the availability of either good reference genomes or an integrated reference transcriptome to which scRNA-seq reads can be mapped to build single cell expression profiles. Given the difficulty of culturing most of the enormous diversity of single-celled eukaryotic life forms, such resources are typically not available. In these cases, an *ad hoc*, partial transcriptomic reference can still be built by assembling *de novo* the whole set of mRNA reads from the total pool of captured cells. In this study, we aim to test if such an approach is feasible. For this purpose, we use full-length mRNA sequencing to explore and analyze expression profiles of the single celled chrysophyte *Ochromonas triangulata*, for which no genome reference is available. Cells from two distinct growth phases in culture were FACS-sorted and single-cell transcriptomic libraries were prepared following the Smart-seq2 protocol, which provides the full-length transcript coverage needed for *de novo* assembly of its partial transcriptome.

Materials and methods

Culture conditions

Clonal, unialgal, non-axenic stock cultures of *O. triangulata* strain RCC21 (Roscoff Culture Collection), were grown in batches of filter-sterilized K/2 medium (Keller et al. 1987, modified by substituting natural seawater for artificial seawater prepared according to Berges et al. 2001). The cultures were serially kept at 18°C under a 12 h photoperiod (120 µmol m^{-2} s^{-1} photon flux measured with a QSL-100 spherical sensor, Biospherical Instruments) by diluting 10 times a fraction of each batch into fresh medium approximately every ten days. Following this routine, representative cells from the fast-growing phase were sampled after 2 days from the refreshing date, whereas representatives of the slow-growing phase were sampled 11 days after refreshing. *O. triangulata* cell abundance prior to sorting was monitored daily using a CytoFLEX flow cytometer (Beckman Coulter). Cell discrimination was based on joint side scatter signal (SSC) and chlorophyll *a* autofluorescence detected in the FL3 channel (EX 488, EM 690/50). Cell abundance of the co-cultured, indigenous bacterial community was based on fixed samples (1% formaldehyde final concentration) kept at 4°C upon fixation and measured within 24 h from sampling after staining with SYBR™ Green I (Invitrogen) and jointly detected in the SSC and FL1 channels (EX 488, EM 525/40).

Sorting

Cells from both growth phases were stained with LysoSensor™ Blue DND-167 (Invitrogen), used as both an indicator of cell viability and presence of food vacuoles. Single cell sorting was performed with a MoFlo Astrios EQ cell sorter (Beckman Coulter) using 355 nm and 640 nm lasers for excitation, 100 µm nozzle, sheath pressure of 25 psi and 0.1 µm sterile filtered 1 x PBS as sheath fluid. Side scatter was used as trigger channel. Sort decisions were based on gates indicating presence of chlorophyll *a* (640-671/30 vs SSC) in combination with LysoSensor detection (355-448/59). Singlets were cytometrically selected based on the height-to-area relationship of the pulse (640-671/30 Height-log vs 640-671/30 Area-log). Individual cells were sorted based on the most stringent single cell sort settings available (single mode, 0.5 drop envelope) and deposited into 384-well Eppendorf twin.tec® PCR plates containing 2 µL of lysis buffer (see Library preparation section for composition). The sample plates were divided into four regions of equal size and cells from either growth phase were distributed in alternating regions. This was made to even out technical variation derived from downstream plate processing. A validation plate serving as an initial test of success was sorted using two columns with cells from each of the growth stages. The first two wells in each validation column contained 10 and 5 cells, respectively, to be used as

positive controls. The plate holder of the sorter was kept at 4ºC at all times. After the sort, the plates were immediately spun down and left on dry ice until storage at -80°C. Flow sorting data was interpreted and displayed using the sorter-associated software Summit v 6.3.1.

Library preparation

All plates were processed according to the Smart-seq2 protocol (Picelli et al. 2014) with slight modifications to comply with small-volume handling robots. The oligo-dT (Smart-dTV30) primer, template switching oligo (TSO) and pre-amplification (IS-PCR) primers were all modified with a 5'-end biotin molecule (Table 1. This is crucial to increase cDNA yield by avoiding concatenation of TSOs after the first stand reaction (Kapteyn et al. 2010, Turchinovich et al. 2014). The lysis buffer contained a final concentration of 0.2 % Triton - X100 (Sigma), 1 U/µL RNAse inhibitor (cat. no. 2313A, Takara), 2 mM dNTPs (ThermoFisher), Smart-dTV30 primer (IDT) and ERCC RNA Spike-in Mix (cat. no. 4456740, ThermoFisher) diluted $4 \cdot 10^5$ times. The lysis buffer was cold-dispensed in 2 µL fractions using the MANTIS liquid dispenser (Formulatrix). Plates with sorted cells in lysis buffer were thawed and cDNA generation was conducted in 5 µL reactions containing a final concentration of 1x Superscript II buffer, 5 mM DTT, 1 M $MgCl_2$, 1 U/µL RNAse inhibitor (cat. no. 2313A, Takara), 5 U/µL SuperScript™ II Reverse Transcriptase (Invitrogen) and 1 µM TSO (Qiagen). The master mix was dispensed using the MANTIS liquid dispenser followed by mixing for 1 min at 1800 rpm on a plate shaker (Biosan). First strand reaction was run at 42°C for 90 min, followed by 10 cycles of 50°C for 2 min and 42°C for 2 min, with a final 5 min extension at 72°. A 72°C initial denaturation step was omitted as this had shown to have no effect on the results.

Table 1. Priming sequences used for the Smart-Seq2 protocol.

oligo	sequence
Smart-dTV30	5'-Biotin-AAGCAGTGGTATCAACGCAGAGTACT30VN-3'
TSO	5'-Biotin-AAGCAGTGGTATCAACGCAGAG-TACATrGrG+G-3'
IS-PCR	5'-Biotin-AAGCAGTGGTATCAACGCAGAGT-3'

Pre-amplification was run in 12.5 µL final volume containing a final concentration of 1x KAPA HiFi HS RM (Roche), and 10 µM IS-PCR primer. PCR reactions were run at 98°C for 3 min, followed by 24 cycles of 98°C for 20 s, 67°C for 15 s and 72°C for 6 min, with a 6 min final extension at 72°C.

Bead cleanup of cDNA was automated using the Biomek NXP liquid handler (Beckman Coulter), AMpure beads and an Alpaqua 384 post magnet with spring cushion technology (Alpaqua). In short, 10 µL beads were added to the 12.5 µL cDNA, the plate was mixed for 1 min at 1800 rpm on a plate shaker and incubated for 8 min before spinning the liquid down and placing the plate on a magnet for 5 min. While on the magnet, the supernatant was removed and 25 µL of freshly prepared 80% EtOH was added and then aspirated after 30 s. The washing was repeated and the plate was left to dry for 2-3 minutes. Elution of cDNA was done by adding 13 µL of water to the wells and the plate was then mixed before 5 min of incubation at RT. Plates were briefly centrifuged and then put on the magnet for 2 min before transfer of 10 µL to a new plate.

Twenty-two single reactions were randomly chosen across the plate for quantification using Qubit HS DNA ready mix, of which 11-15 samples were also run on the Bioanalyzer using the HS DNA chip (Agilent).

Sequencing

Nextera XT library preparations were run in 5 µL reactions where all reagent volumes were scaled down ten times. The purified cDNA was diluted to approximately 200 pg/µL based on the average cDNA concentration of the 22 randomly chosen reactions. The ATM and TD reagents were pre-mixed and distributed in a 384-well plate before adding cDNA using the Mosquito HV contact dispenser (SPT Labtech). Barcodes of the Illumina Nextera™ DNA UD index set A-D (Illumina) were diluted 1:5 prior to the library amplification. Reactions were run for 12 cycles according to cycling conditions recommended by the manufacturer. For all steps including manual mixing, a plate shaker run at 1800 rpm and a plate spinner were used. Single reactions from each plate were pooled according to their originating plate and purified using 1:0.8 AMPure XP beads. The two pools from each plate were run on separate lanes on a NovaSeq 6000 SP v1.5, PE 1 x 150 bp, including 1% PhiX spike-in. All sequence data generated in this project have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI and made publicly available under accession number PRJEB60973.

De novo *transcriptome assembly and read mapping*

Trinity (Grabherr et al. 2011) was used for *de novo* combined assembly of the *O. triangulata* partial transcriptome (reference). Reads from all the individual cells were combined for the task. Following assembly, reads from individual cells were aligned to the reference transcriptome using Bowtie (Langmead et al. 2009) and transcript abundance was estimated using RSEM (Li and Dewey 2011).

Clustering and transcript analysis

Before clustering, data from cells with very low read coverage (fewer than 50 000 reads) were removed. Additionally, contigs that were observed in only a single cell or showed a total of over 20 million TPM (Transcripts Per Million) across the population were filtered out to avoid under- and over-representation bias, respectively. The transcript count matrix was dimensionally reduced using *t*-SNE (*t*-distributed Stochastic Neighbourhood Embedding). Further, clustering was performed using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm (5 minimum points within $\varepsilon = 4.5$). Transcript accumulation and transcript coverage curves following rarefaction were generated using the iNEXT R package (Hsieh et al. 2016) analogously to procedures used for obtaining curves of species accumulation and species coverage (Chao et al. 2014), using as counterparts transcript identity for species identity and cell identity for sample site.

Transcriptome annotation

The *de novo* assembled transcriptome was annotated using blastx as implemented in the BLAST+ suite (Camacho et al. 2009). UniProtKB/Swiss-Prot was used as the database for the annotation. MGkit (Rubino et al. 2014), AGAT (v1.1.0, Dainat et al. 2021) and TransDecoder (v5.5.0, Haas 2018) were used to obtain only the annotated parts of the assembled contigs. The longest isoforms were selected using a custom python script. Reads from the individual cells were re-mapped to the annotated transcriptome for further expression analysis.

Differential expression analysis

Pairwise differential expression analysis was carried out using DESeq2 (Love et al. 2014). The analysis was performed with reads mapped to both the assembled transcriptome and annotated transcriptome separately. Transcripts with a Benjamini-Hochberg adjusted p-value less than 0.05 were considered as differentially expressed.

Metabolic mapping

The expression matrix of annotated transcripts was mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database. Pathview (Luo and Brouwer 2013) was used to map mean expression levels of proteins in the three clusters to metabolic maps or pathways obtained from the KEGG Orthology database. To reduce the number of maps to a tractable number, only maps with at least 1 matched protein, 0.25 match ratio (number of matched proteins/ number of proteins in pathway), and 0.1 differential expression ratio (number of differentially expressed proteins/ number of matched proteins) were retained. The remaining pathways were manually inspected and pathways where

the only differentially expressed genes were not specific to the pathway were removed.

Detection and annotation of rRNA reads

Detection of sequence reads originating from rRNA was carried out with RiboDetector (Deng et al. 2022). For taxonomic classification of putative rRNA sequences, reads of 150 nucleotides or longer and represented at least twice in the whole dataset were annotated using the SINTAX algorithm implementation in VSEARCH (Edgar 2016, Rognes et al. 2016) against the PR2 Reference Sequence Database version 4.14.0 (Guillou et al. 2013) for eukaryote identification and against the SILVA 138 SSU Ref NR 99 (Yilmaz et al. 2014) for prokaryotic community identification.

Results

Flow cytometry and sequencing

For our investigation, the relatively small (~8 µm in cell diameter) mixotrophic chrysophyte *Ochromonas triangulata* was used as model microeukaryote. Cells were sampled from two distinct growth stages during its life cycle in batch culture (Figure 1A), either at the early onset of stationary phase (batch B14 in Figure 1A) or at the onset of the exponential growth phase (B15 in Figure 1A). Cells from B15 were steadily growing at a rate of 0.6 doublings per day at the time of sampling and are hereon referred to as fast-growing cells. In contrast, cells from B14 were growing at a rate linearly declining by a factor of 0.08 (0.2 doublings per day at the time of sampling) and are referred to as slow-growing cells in the rest of the text. The transition from fast to slow growth occurred around day 5-6 after inoculation and was associated with a decrease in the abundance of the fast-growing bacteria in co-culture with *O. triangulata* (Supplementary Figure 1), which transitioned into a persistent community concurrently with *O. triangulata* transition into linear growth.

Eukaryote cells were sorted based on the joint signal of chlorophyll *a*, intrinsic to *O. trinagulata*, and LysoSensor Blue, a food vacuole reporter (Carvalho and Granéli 2006) used here to discriminate living from senescent cells. Cytometric analysis of the sample already showed a slightly increased chlorophyll *a* fluorescence in the fast-growing cells compared to the slow-growing (Figure 1B). The latter group also featured a small number of cells of reduced LysoSensor signal. These putatively non-feeding cells were assumed to be senescent based on the concomitant signal of LysoSensor and fluorescein diacetate (FDA), or lack thereof, after a viability assay (Supplementary Figure 2). Therefore, they were excluded from the sorting selection.

After sorting, single-cell mRNA libraries were prepared following the SmartSeq2 protocol, which preserves the full-length transcript needed for *de*

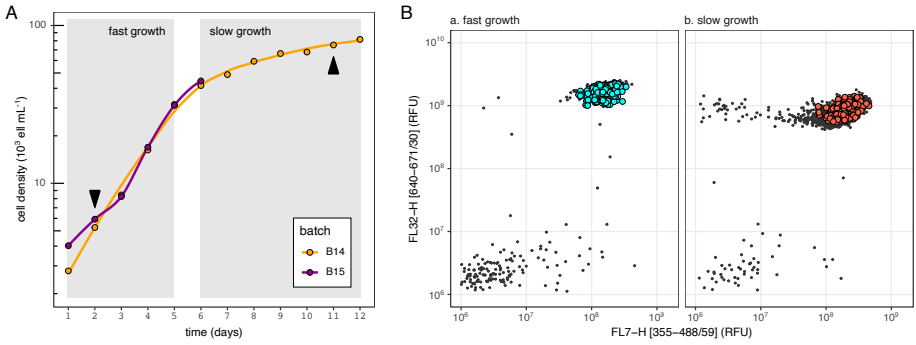


Figure 1. Sampling Ochromonas triangulata. (A) Growth curves of source cultures for cells during exponential growth stage (fast growth) and linear growth stage (slow growth). Batch labels refer to the source culture for cells in slow (B14) and fast growth (B15). Arrowheads indicate the sampling points for each stage. (B) Flow cytograms showing placement of FACS-sorted O. triangulata cells (colored dots) for each growth stage.

novo transcriptome assembly. A total of 744 single-cell transcriptome libraries were sequenced with the Illumina NovaSeq 6000 instrument extending over two SP lanes, each loaded with 378 and 366 cells, and yielded median sequencing depth of 1.4 and 1.5 million reads per cell, respectively.

Transcriptome assembly and clustering

De novo assembly of filtered and quality trimmed reads pooled across all cells produced a total transcriptome of ~166 000 transcripts (median length 432 bases; contig N50 1 358 bases). Such a size was deemed unrealistic for a small sized unicellular phytoflagellate (Liu et al. 2017). Accordingly, a requirement of presence in at least 2 cells was set as a threshold to accept a transcript in the assembly. This reduced the transcriptome size to 60 307 transcripts (media length 835 bases), which might still represent an inflated number compared to the ~20 000 transcripts reported for other *Ochromonas* strains (Lie et al. 2018). Stricter recovery thresholds produced smaller transcriptome sizes asymptotically (Supplementary Figure 3), but this relationship never approached the asymptote within a reasonable minimum. Lacking a reference, the second most permissive threshold was chosen to preserve the most comprehensive transcriptome. In addition, exclusion of cells with very low read coverage (carrying less than 50 000 reads per cell) reduced the number of cells under scrutiny to 721. The dataset resulting from applying both thresholds was chosen as the basis for read mapping, generating the transcript expression table supporting all subsequent analysis.

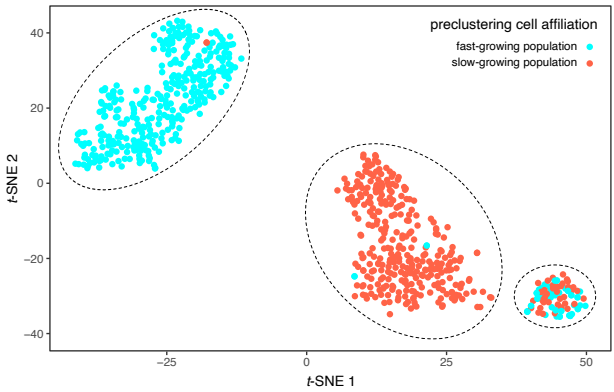


Figure 2. *t*-SNE visualization of *Ochromonas triangulata* expression clusters. Each dot represents one cell color-coded based on its sampling origin – either the fast-growing (cyan) or slow-growing (red) cell consortia. Clusters identified by DBSCAN are enclosed by a dotted line.

DBSCAN clustered the cells based on dimensionally reduced transcript expression and confirmed that the expression profiles of each respective cell consortia were distinct and well resolved depending on growth stage (visualized after *t*-SNE dimensionality reduction in Figure 2). The clustering algorithm identified 332 cells as fast-growing and 314 as slow-growing, almost all of them mapping consistently to their sampled origin, with only three individuals mapping in disagreement. Interestingly, the expression-based cell clustering also revealed a smaller but discrete third cell group, evenly made up of individuals cells from both the fast (37 cells) and slow (38 cells) growth stages. The presence of an additional cell type in either of the sampled cultures was completely unexpected, and thus a priori information about its character was lacking. Therefore, this stage is hereafter referred to as cryptic and the identity of each of these cells was reclassified accordingly. Consequently, in the following text, preclustering affiliation refers to cell identity as determined by sorting conditions, whereas postclustering affiliation refers to cell identity as demarcated after DBSCAN clustering.

Characterization of transcriptomic sampling effort

Regarding transcript richness per group, the distributions of the number of unique transcripts per cell for the fast and slow growth stages were overlapping (Figure 3A). While median transcript richness per cell of the fast-growing consortium was slightly higher than that of the slow-growing, both medians were at or above 1000 different transcripts recovered per cell, regardless of pre- or postclustering affiliation. The small differences observed in the distributions when considering reclassification after clustering are related to the resolution of the cryptic group, which is associated with an ostensible reduction in unique transcripts per cell (median of 628) and a distribution that barely

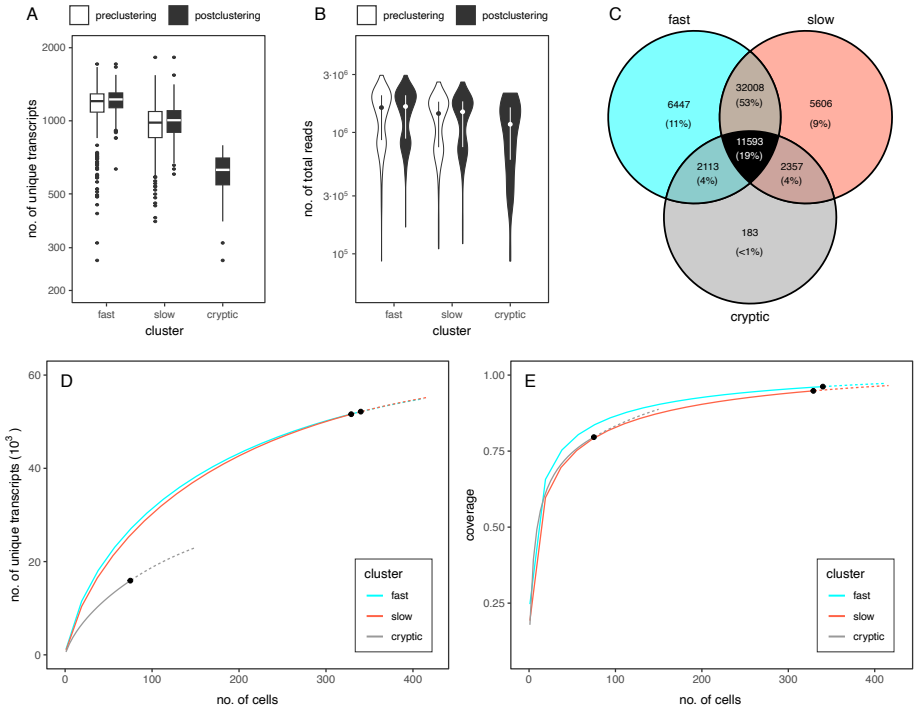


Figure 3. Completeness analysis per group. (A) Distributions of transcript richness per cell. (B) Distributions of read counts per cell. Both A and B show data separated by pre- and postclustering affiliation (C) Venn diagram showing the number of shared and exclusive transcript sets among postclustering groups (D) Transcript accumulation plot, and (E) partial transcriptome coverage plot for each postclustering group. In (D) and (E), the black dots indicate sample size, the solid lines follow sample rarefaction and the dotted lines represent an extrapolation based on the rarefaction model for each group.

overlapped any of the other two groups. This decrease in the number of unique transcripts is still not associated with a comparable decrease in read support for cells in the cryptic group (Figure 3B), suggesting that such a reduced transcript representation per cell was not an artifact originating from poor sampling depth, but rather a consequence of a real biological phenomenon.

In total, almost 20% of the transcripts were shared between the three expression clusters (Figure 3C). However, almost three quarters of the whole transcriptome (72%) were shared between the fast and slow-growing groups, and slightly more than half (53%) are exclusive to these two clusters. In contrast, only 4% of the assembled transcripts are shared between the cryptic and either of the other two clusters, and less than 1% are unique to this cluster.

To further explore the degree of transcriptome recovery within each cluster, a rarefaction analysis was performed (Figures 3D-E). In terms of transcript identity, rarefaction curves for the fast and slow growth groups showed a

marked asymptotic behavior before reaching cell sampling depth, which suggests that the number of sampled individuals from each consortium was sufficient to convey a large proportion of the total transcript diversity (Figure 3D). This behavior, to a lesser extent, was also apparent in the rarefaction curve for the cryptic cluster. Recovery of transcripts unique to the cryptic group was, however, observed to be erratic: none of the transcripts in this fraction (grey colored in Figure 3C) were recovered in more than 3 cells. The reduced slope of the curve compared to the other two denoted a lower transcriptome richness characteristic of the cryptic group. The extrapolation of its rarefaction model points towards a 40% reduction compared to either the fast or slow-growing stages, which, in turn, displayed similar transcript richness. At the same time, all expression clusters showed coverage of stage-specific transcriptomes above 75%, with the fast and slow cell clusters with higher number of representatives very close to full transcriptome completeness (Figure 3E).

Differential expression of annotated transcripts

To understand the biological features driving the underlying expression structure, transcripts were annotated against the UniProt database. Notably, only 30% of all transcripts could be assigned biological meaning (Supplementary Figure 4), highlighting the problem of underrepresentation of protist sequences in publicly available annotation databases (Keeling et al. 2014). The annotated set, comprising a total of 17 883 transcripts, was then used to perform differential expression analysis with the aim to understand the distinct biological character of each stage.

Pairwise comparisons among the three distinct cell clusters revealed 538 differentially expressed annotated genes (DEGs). To expose integrated biological information conveyed by the set of DEGs, the KEGG Pathway database was queried using transcript identity from UniProt annotation. The relevant pathways that resulted from the search are collected in Figure 5. Within these, four main pathway families can be recognized: those associated to transcription and translation processes, those associated to vesicle maintenance and membrane trafficking, those related to photosynthetic activity, and the most numerous associated to general cell metabolism. DEGs in the KEGG pathways detected from any pairwise comparison against the cryptic group are, in all instances and without exception, downregulated in this stage.

Three pathways related to transcription and translation (ribosome, protein processing and proteasome) collect the highest number of differentially expressed genes. In addition, two other key components in the progression from transcript to functional protein (spliceosome, protein export) also present a small number of differentially expressed genes. No clear pattern in differential gene expression between the fast and slow growth groups is apparent in pathways from this class, with only a few genes (maximum 16 in the ribosome

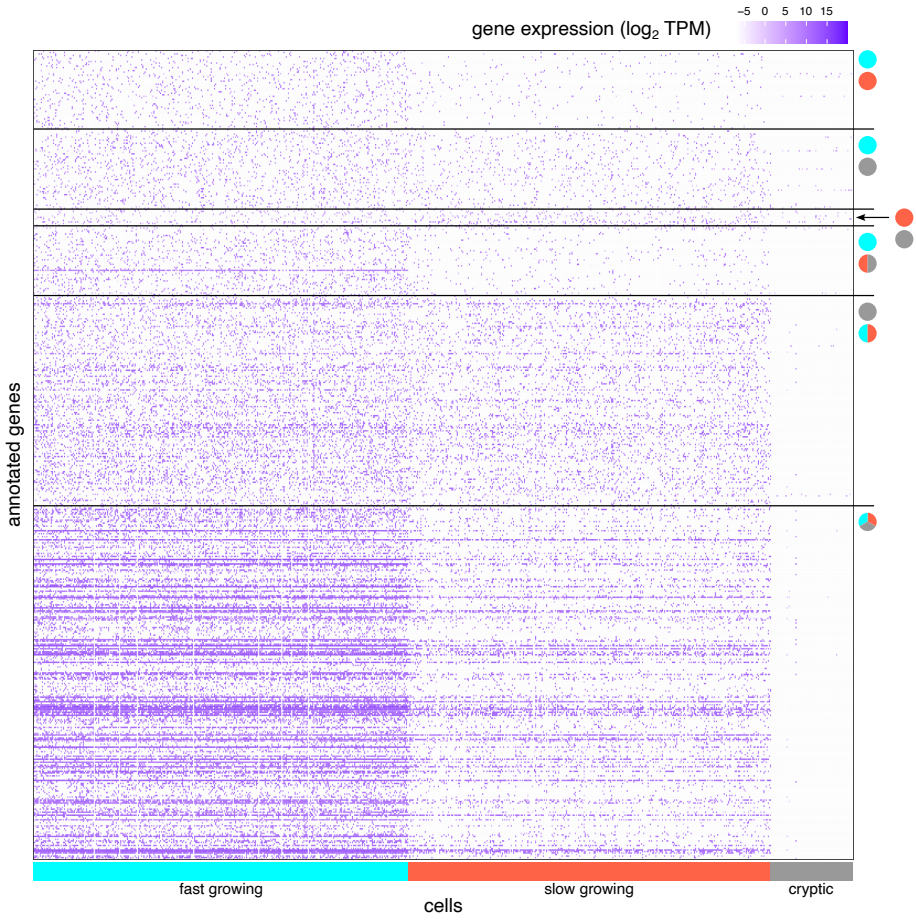


Figure 4. Heatmap of annotated genes that were differentially expressed between different *Ochromonas triangulata* growth stages. Colored rectangles in the horizontal axis indicate cell affiliation to clustering group. Circle pairs indicate pairwise comparisons between stages. Those cases where gene expression in one group is significantly different to that of the other two groups, the joint is represented by a split circle in which the color of each half encodes the identity of each member. A three colored circle represents cases in which gene expression differs significantly for any possible pairwise comparison. The colors in circles and rectangles correspond to fast-growing (cyan), slow-growing (red) and cryptic (grey) stages.

protein set, Figure 5) showing significantly different expression levels. In contrast, downregulation of genes in this class associated to the cryptic group is prevalent (28 and 33 genes when comparing it to the slow-growing or fast-growing groups, respectively). This agrees with the pattern observed in Figure 4 and suggests reduced transcriptional activity in the cryptic stage when compared to the other two.

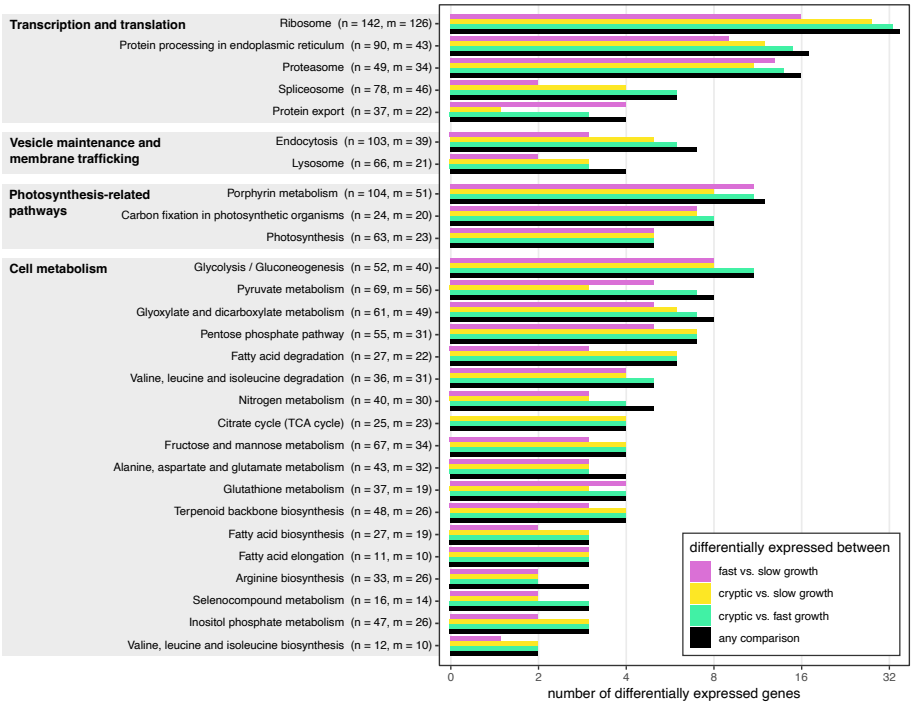
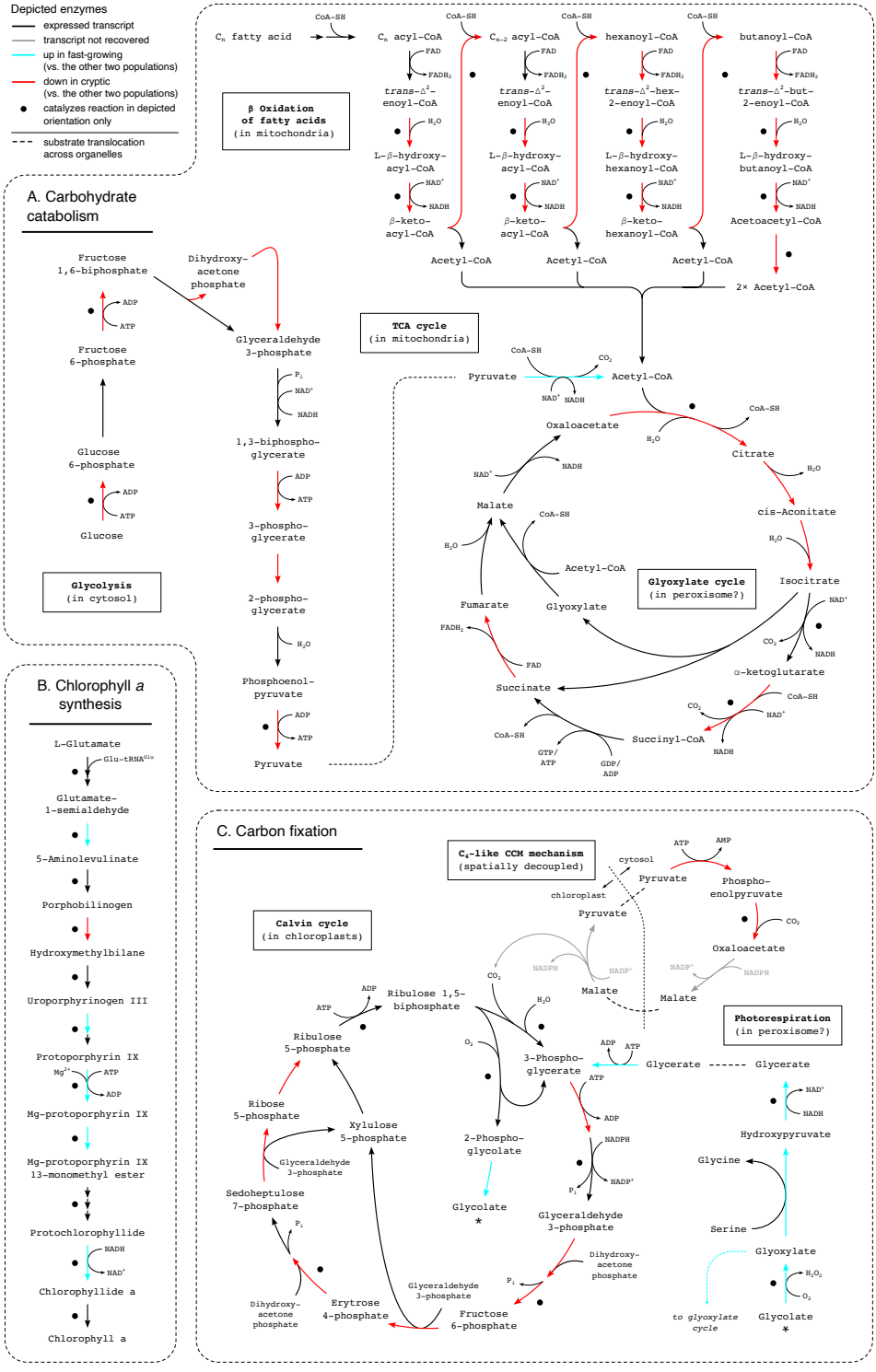


Figure 5. Summary of KEGG pathways associated to multiple differentially expressed genes. For each pathway label, *n* indicates the total number of genes that belong to the pathway and *m* the number of genes for which expression was recovered in our transcriptomes.

The second class (endocytosis and lysosome pathways) provides some insights into phagocytotic behavior, revealing joint downregulation of clathrin and AP-2 (both essential coating components of vesicles resulting from endocytosis) in the cryptic group as well as high expression of lysosomal proteases (homologs of tripeptidyl-peptidase 1 and cathepsin A, B, D, F and X) in both fast and slow-growing expression clusters in parallel with negligible expression of these proteases in the cryptic stage. These observations, although limited, provide evidence of active lysosomal digestion in the growing stages, which in turn seems to be absent from the cryptic stage.

Finally, the third and fourth pathway groups, represented by the pathways associated to photosynthesis and carbohydrate metabolism, harbors the highest number of DEGs featured in the whole set of KEGG and provides the clearest picture associated to physiological differences between cell stages Figure 6). In all cases, DEGs associated to this class showed either significantly higher expression in the fast-growing stage, intermediate in the slow-growing and lowest in the cryptic (cyan arrows in Figure 6), or downregulation in the cryptic stage with no significant differences between the other two groups (red arrows in Figure 6).


Depicted enzymes
expressed transcript
transcript not recovered
up in fast-growing (vs. the other two populations)
down in cryptic (vs. the other two populations)
catalyzes reaction in depicted orientation only
substrate translocation across organelles
β Oxidation of fatty acids (in mitochondria)
A. Carbohydrate catabolism
Glycolysis (in cytosol)
TCA cycle (in mitochondria)
Glyoxylate cycle (in peroxisome?)
B. Chlorophyll a synthesis
C. Carbon fixation
Calvin cycle (in chloroplasts)
C4-like CCM mechanism (spatially decoupled)
Photorespiration (in peroxisome?)

Widespread down-regulation of carbohydrate catabolism is a pronounced feature of the cryptic stage, apparent through the downregulation of key enzymes of the glycolysis pathway (phosphofructokinase-1 and pyruvate kinase), the β oxidation of fatty acids (enoyl-Coa hydratase, β-hydroxyacyl-CoA dehydrogenase and acyl-Coa acetyltransferase) and the tricarboxylic acid (TCA) cycle (citrate synthase) (Figure 6A). It is important to note that the expression of all enzymes involved in β oxidation and most enzymes in the TCA cycle is recovered in all groups, including the cryptic stage. Hence, low expression levels in this group compared to the other two suggests low levels of energy production for cell maintenance.

In contrast, observable in the fourth KEGG pathway with highest number of DEGs (porphyrin metabolism, Figure 5), exacerbation of chlorophyll *a* synthesis was manifested in the fast-growing stage (Figure 6B). It involves the upregulation of 5 enzymes in the pathway, including magnesium chelatase, which catalyzes the first committed step of chlorophyll *a* production, and protochlorophyllide reductase, which generates chlorophyllide *a*, the immediate precursor of chlorophyll *a*. A putative increased chlorophyll *a* production in the fast-growing stage is in accordance with increased chlorophyll fluorescence signal in the sorting cytograms (Figure 1B).

Finally, downregulated key enzymes of the Calvin cycle in cryptic cells (fructose 1,6-biphosphatase and fructose-bisphosphate aldolase, both catalyzing unidirectional reactions, Figure 6C) indicates little or no carbon fixation in this group. At the same time, upregulation of photorespiration enzymes in the fast-growing stage suggests high CO_2 fixation activity associated to these cells, since this is a compensatory mechanism to counteract the efficiency loss in carbon fixation due to the oxidase activity of rubisco. Photorespiration is expected to occur in peroxisomes (Sandalio and Romero-Puertas 2015), which are not characterized in chrysophytes but are reported in other stramenopiles both based on microscopic and genomic evidence (Gabaldón 2010, Mix et al. 2018). This, in combination with increased chlorophyll *a* production, points towards growth conditions mostly supported by photosynthesis.

Among DEGs not associated to a relevant KEGG Pathway, only three show significant upregulation in the cryptic group. Based on UniProt annotation, these are identified as a chromosome partition protein Smc (UniProt A9BZW2), a nitrite reductase (P42435) and a formate dehydrogenase

Figure 6. *In previous page.* Differential expression in key metabolic pathways. Colored arrows indicate: enzymes present (black) or missing (grey) in the total transcriptome, enzymes showing significantly higher expression in the fast-growing stage, intermediate in the slow-growing stage and lowest in the cryptic stage (cyan) and enzymes downregulated in the cryptic stage with no significant differences between the other two groups (red). Dashed lines indicate substrate translocation across organelle boundaries. Black dots indicate enzymes that only catalyze the reactions in the depicted direction.

(Q795Y4). All these hits correspond to prokaryotic entries in the database, although the three genes have also eukaryotic homologs (e.g. Smc2 in *Saccharomyces cerevisiae* [P38989], nitrite reductase in *Giardia intestinalis* [E2RTZ4] and a formate dehydrogenase in *Candida boidinii* [O13437]). Interestingly, a similar phenomenon is observed among the genes unique to the cryptic cluster that could be successfully annotated (42 out of 183). None of these, despite being unique to the group, are among the identified DEGs, most likely because their characteristic erratic recovery causes their expression to lack statistical strength. Nonetheless, they are all associated with prokaryotic hits in the UniProt database. Of these 42 genes, 22 lack any eukaryotic homolog in the database, and therefore 10% of the genes (22 out of 183) unique to the cryptic stage could be in fact prokaryotic. Further exploration of the dataset is needed to gain insights into this phenomenon, but possible explanations are considered in the discussion section.

Taxonomic annotation of rRNA carry-over reads

Ribosomal RNA (rRNA) makes up approximately 80 to 90% of the total RNA content of a single cell (Palazzo and Lee 2015). Therefore, even though most scRNA-seq protocols take advantage of mRNA polyadenylation to enrich for mRNA in scRNA-seq libraries, rRNA leakage is unavoidable (Ku and Sebé-Pedrós 2019). This is an undesired nuisance in most studies where the taxonomic affiliation of the organism under investigation is known. However, if the aim is to target an environmental population for which taxonomic information is lacking, the presence of rRNA can be a great resource to classify taxonomically the transcriptomes resulting from a scRNA-seq effort. We tested the feasibility of this approach with our dataset by attempting to directly annotate putative rRNA reads identified by RiboDetector.

Overall, 241.7 M reads were predicted to be of rRNA origin across all samples (24% of the total), although rRNA read coverage was uneven across cell types, constituting 33%, 16% and 4% of the fast-growing, slow-growing and cryptic cells, respectively (Supplementary Figure 5). After discarding singletons and reads shorter than 150 nucleotides, the resulting set was successfully dereplicated into 4.5 million unique sequences. Of these, one third (33%) could be taxonomically classified with confidence (>0.8 bootstrap support).

Despite the unsuccessful classification of 67% of the sequences, annotated read recovery across samples was high (median 106 688 sequences per sample) and sufficient to correctly annotate 88% of the cells (653) as *Ochromonas triangulata* (median 23 811 reads per sample), with an average of 8.5% of total reads per cell assigned to the best hit.

On the other hand, taxonomic annotation of the 91 remaining cells was poor even though the number of annotated rRNA reads (median 17 265 reads per cell) was still apparently high. The annotation of these cells was dominated by spurious taxa: 70% (57) were annotated as land plants (44% *Oryza*, 11%

Epifagus, both based on chloroplast rRNA, and 3% *Cucumis*), 25% were annotated as fungi (20 samples, all *Candida* except one sample with only 14 reads, annotated *Biatoropsis*) and 3% were annotated as Chlorophyta (3 samples, as *Prototheca*). Taxonomic annotations of the remaining cells (11) were diverse and each designated at most two cells. In all cases, less than 500 reads (median 81) supported the top assignation, which in turn accounted for on average only 1.6% of all rRNA reads per sample. Further stressing the indiscriminate annotation of these group, only 5 of these cells featured reads correctly annotated as *O. triangulata*, but these were always below 1% of the total.

Regarding taxonomic affiliation across expression clusters, 16 of the 91 cells that couldn't be robustly annotated lacked postclustering affiliation. These cells were excluded from the transcriptomic analysis because they failed in mapping at least 50 000 reads onto the assembled transcriptome (see Cluster analysis section). Interestingly, the remaining 75 cells that could not be annotated belonged, without exception, to the cryptic stage. Median rRNA abundance in the cryptic stage is low (approx. 43 000 raw reads per cell) which resulted in less than one fifth of unique sequences (20 083, median value) that enabled successful annotation of all the other cells.

A generally reduced expression landscape, which also includes lower demand for rRNA (as suggested by generalized downregulation of ribosomal proteins, Figure 5) might explain insufficient rRNA levels for unambiguous annotation of cells characterized by low transcriptomic activity.

The 16S rRNA read recovery across the samples was relatively low (median 2 807 reads per sample) but saturated (Supplementary Figure 6) and sufficient to speculatively characterize the prokaryotic community composition associated with individual *Ochromonas triangulata* cells.

The prokaryotic communities of the fast-growing and cryptic cells were internally uniform (Figures 7A and 7B; median Bray-Curtis dissimilarities 0.1745 and 0.1498 within the groups, respectively) compared to the slow-growing cells (median Bray-Curtis dissimilarity 0.2891 within the group; Figure 7B). The prokaryotic communities were significantly different between the fast- and slow-growing (PERMANOVA R^2=0.4521 P=0.001; median Bray-Curtis dissimilarity 0.4340), fast-growing and cryptic (PERMANOVA R^2=0.7786 P=0.001; median Bray-Curtis dissimilarity 0.797) and slow-growing and cryptic cells (PERMANOVA R^2=0.3127 P=0.001; median Bray-Curtis dissimilarity 0.4338). The alpha diversities between the communities were not significantly different (Supplementary Figure 6C).

The proportion of the 16S rRNA to all rRNA was significantly higher in the cryptic cell cluster compared to the fast- and slow-growing stages (Wilcoxon $P < 2.2 \cdot 10^{-16}$ and $P < 0.009$, respectively) whereas the proportion of 16S rRNA to all RNA was not significantly different between the groups (Figure 7C). This is in contrast with the proportion of the 18S rRNA to all rRNA

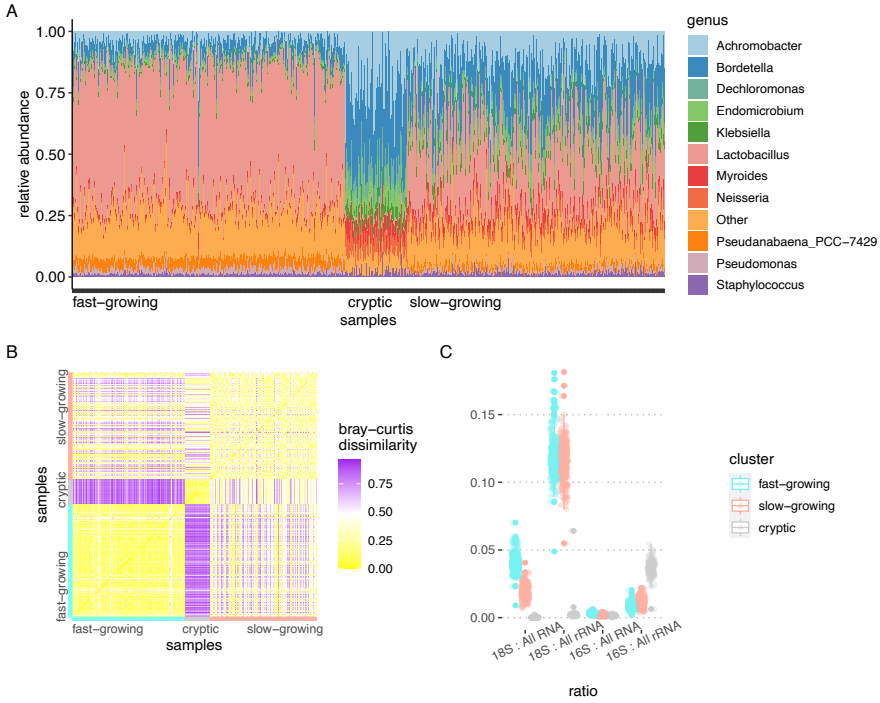


Figure 7. The prokaryota associated with the *O. triangulata* cultures. (A) Relative abundance of the eleven most abundant prokaryotic genera associated with the different expression clusters. (B) Bray-Curtis dissimilarity of each sample' associated prokaryota with the rest of the samples. (C) Ratios of 18S and 16S rRNA to the rest of the RNA or rRNA obtained from the samples.

and all RNA reads, where the 18S rRNA proportion for the cryptic group is significantly below the other (Wilcoxon test $P < 2.2 \cdot 10^{-16}$ and $P < 2.2 \cdot 10^{-16}$, respectively). The differentially abundant families (determined by a consensus of DESeq2, LinDA, Maaslin2, ANCOMBC, and ALDEx2; adjusted $P < 0.05$; Supplementary Figure 6C) between the cell clusters include *Flavobacteriaceae*, *Endomicrobiaceae*, *Alcaligenaceae*, *Burkholderiaceae*, *Neisseriaceae*, *Rhodocyclaceae*, *Enterobacteriaceae* (lowest abundance in the fast-growing population, highest abundance in the cryptic population) and *Lactobacillaceae*, *Streptococcaceae*, *Pseudanabaenaceae*, *Rhizobiaceae*, *Anaplasmataceae*, *Pseudomonadaceae* and *Staphylococcaceae* (lowest abundance in the cryptic population, highest abundance in the fast-growing population). Interestingly, the slow-growing stage was always the middle one among the differentially abundant families.

Discussion

Our results show that transcriptomic profiling of *O. triangulata*, a small microbial eukaryote, at single cell resolution is technically feasible, even in the absence of a reference genome. The SmartSeq2 protocol, used to generate the mRNA libraries of FACS sorted single cells of *O. triangulata*, provided sufficient output while preserving the full-length transcript information necessary to generate a working transcriptome assembled *ad hoc*. Using this transcriptome as the basis for the analysis of gene expression, we were able to discern three distinct cell clusters of *O. triangulata* originating from two contrasting growth stages, therefore exposing a group of cells that was not expected *a priori*.

Transcript recovery for each of these three groups was satisfactory. Concerns about the need of very large sample sizes to discern expression patterns in microbial eukaryotes have been raised previously (Liu et al. 2017), indicating that sampling sizes well above 100 individuals would be necessary. This concern is grounded on the joint effect of the low numbers of mRNA molecules coexisting in a small microbial eukaryote (Liu et al. 2017) together with the widespread occurrence of gene transcription in successive bursts (Suter et al. 2011). Indeed, our results show that transcriptome recovery at cell sampling depth (represented by black dots in Figure 3D and 3E) is far from saturation in all three groups. Underlying Figure 3D are a notable proportion of transcripts that were recovered in only 1 cell (16%, 26% and 28% for the fast growing, slow growing and cryptic cell clusters, respectively). This somewhat erratic transcript recovery per cell could be an inevitable consequence of transcriptional bursts. This phenomenon, however, might not be determinant to reveal expression clusters. Figure 3E shows that, at least for *O. triangulata*, 100 cells would have been sufficient to recover ~80% of the transcriptome in each of the three groups (50 cells in the case of the fast-growing populations, which shows the highest mRNA recovery, Figures 3A and 3B). We therefore suggest that sampling in the order of 100 individuals might be enough to characterize microeukaryote expression clusters based on single cell expression profile.

While we found no technical limitations in deploying scRNA-seq, annotation issues are still pervasive. In our dataset, two thirds of the comprehensive transcriptome remained unannotated (Suppl. Figure 4). Consequently, of all the data generated in our experiment, only approximately 11% of the reads contributed to characterize the expression levels of mRNA transcripts with putative biological function (Suppl. Figure 5). On the other hand, the dataset is characterized by a large proportion of dark expression (i.e. reads that mapped to transcripts devoid of known function, 45% in our dataset), and therefore much biological significance remains unavailable due to weaknesses of annotation. Unavoidably, organisms that are poorly represented in

annotation databases (as is the case for most non-model microeukaryotes) will be susceptible to this issue.

In our dataset, this is well exemplified by the cryptic cell cluster. This cell group was characterized by a lower number of well supported unique transcripts which accounted for a good representation of the expression landscape of the group (covering approx. 80% of its full transcriptome). In turn, most of these transcripts were shared with one or both of the other two groups, while only 1% were unique to this group.

The question of what these cryptic cells represent remains open. We speculate about the possibility that these are cells on their way towards encystment. The formation of endogenous cysts, known as stomatocysts, is a well-known characteristic phenomenon of chrysophytes (Kristiansen and Škaloud 2017). Stomatocysts are spherical, ellipsoidal, or ovate silica structures enclosing the protoplast and bearing a single opening that is closed upon cyst maturity (Hibberd 1977). Several *Ochromonas* species are known to be capable of stomatocyst formation in culture (Hibberd 1977, Andersen 1982). Interestingly, stomatocyst formation is mediated by the activity of silica deposition vesicles (SDVs) which build up the silica wall while enclosing it. This mechanism could resemble that of frustule-building in diatoms, mediated by acidic SDVs. A low pH within SDVs could explain that cells going into encystment are LysoSensor positive, regardless of their feeding behavior. However, key proteins during frustule formation in diatoms, such as silaffins, cingulins and silacidin (Tesson et al. 2017) are absent in the transcriptome. In any case, no transcriptomic clues are available for the encystment process, and further experiments would need to be performed to provide support for this explanation.

Alternatively, a scenario in which a eukaryotic cell is being colonized or invaded by a bacterial community would we consistent with the prokaryotic nature of at least a portion of the mRNA transcripts unique to the cryptic group. This possibility would also be in accordance with the very erratic recovery associated with these transcripts, since the recovery of non-polyadenylated prokaryotic mRNA, likely caused by random mRNA capture in a similar mechanism driving rRNA leakage, would be expectedly inconsistent. In parallel, if these cryptic cells would truly represent a clearly differentiated stage, a fair proportion of the total transcript pool could be expected to be shared among the three clusters (reflecting the expression of housekeeping genes) but slim pairwise overlaps between the cryptic cluster and the other two groups would be more likely, especially if the cells in the cryptic cluster would depart (e.g. via encystment) from an active state, which is characteristic of the other two. Instead, the relatively high proportion of transcripts that are shared between the cryptic and either of the other two groups (slightly above 2 000 transcripts in both cases, Figure 3C) suggests the bacterial colonization of cells originating from either of the groups.

In any case, this observation indicates the probable presence of reads of prokaryotic origin in the assembly. In our analysis, based on the *de novo*

assembly of a working transcriptome built from the total read pool, we cannot rely on read mapping to a reference genome to sort out the bacterial reads. As a result, all these sequences, generated from non-polyadenylated RNA fragments, end up forming part of the assembly, which can in turn partially explain the inflated size of the transcriptome. The implementation of a *k*-mer analysis algorithm prior to assembly (Wood and Salzberg 2014) could contribute to ameliorate this issue, and could contribute decisively to, for example, elucidate the identity of the cryptic group.

Despite this limitation, differential expression of genes involved in quasi-universal cellular pathways, and therefore well represented in annotation databases, provided insight into the functional state of *O. triangulata* beyond the cryptic group. For example, enhanced photosynthetic activity in combination with seemingly increased CO_2 fixation were characteristic of cells in fast exponential growth. This, together with a concurrent peak in bacterial abundance during this phase seemed to indicate that a boost in *O. triangulata* growth could be enabled by its heterotrophic activity. When prey declined into a stable, potentially predation-resistant community, *O. triangulata* growth transitioned into a slower growth rate driven mostly by photosynthesis. This would be in line with what has been shown before for *Ochromonas* isolate CCMP1393 (Lie et al. 2018, Wilken et al. 2020), and would delineate *O. triangulata* as a mainly phototrophic constitutive mixotroph that could benefit from prey uptake to accelerate growth (group C mixotroph *sensu* Jones 1997).

Another additional source of useful information that was featured in our dataset came from rRNA annotation. Although Smart-Seq2 targets specifically mRNA, the cellular amount of rRNA in a eukaryotic cell is vast enough to allow the leakage of rRNA reads into the dataset. This phenomenon, typically regarded as a nuisance, can be used to the researcher's advantage when taxonomic affiliation of a cell is unknown, as would be the case from a natural sample. In our dataset, about one quarter of the total read volume corresponded to rRNA reads, although it was not equally distributed across cell clusters. This can become limiting when read coverage per cell is relatively low, since we needed >100 000 sequences to correctly annotate *O. triangulata* cells based on direct annotation of rRNA reads. The reason for this might be that the taxonomical information conveyed by these sequences is generally poor because they can originate anywhere in the rRNA operon, much of which is phylogenetically uninformative. In the case of the cryptic group, where only 4% of reads overall originated from rRNA and read support falls one order of magnitude below, annotation becomes misleading. The potential prokaryotic nature of many reads associated to the cryptic cell group could also explain the difficulty in annotating properly the members of this group. Nevertheless, when read support is high, cells could be taxonomically annotated with confidence.

The carry-over of 16S rRNA reads in the sequencing data permits profiling the bacterial communities associated with individual *O. triangulata* cells.

Surprisingly, the community compositions were consistent within the clusters but variable between them despite the similarity of the co-cultured bacterial communities. The internally most consistent bacterial communities were amidst the fast-growing and cryptic cell groups whereas more variability was observed within the slow-growing group. Interestingly, the bacterial community composition of the members of the cryptic group did not depend on their origin in the fast- or slow-growing cells. Also, the slight bimodality in the alpha diversity of the cryptic cells had no connection to this.

We speculate that the differences in the bacterial communities between the groups could be explained by grazing behavior of the *O. triangulata* cells. Initially, low abundance of *O. triangulata* cells would allow the bacteria to grow fast under little grazing pressure. With time and consequently increasing numbers of *O. triangulata*, the peak density of the bacterial community is consumed and the community transitions to a grazing-resistant steady state (Supplementary Figure 1) forcing *O. triangulata* to progress from a fast-growing to slow-growing state. The increased grazing pressure due to higher numbers of *O. triangulata* could selectively affect the community structure of the co-cultured bacteria, which in turn could result in lower stability as well as a shift from the bacterial community associated with the fast-growing population. However, the current study lacks community profiling of the co-cultured bacterial cells and therefore we are not able to confirm the similarity of the individual *O. triangulata* bacterial communities to the bulk bacterial community structure of the fast- and slow-growing culture stages.

The specific bacterial community composition within the cryptic cell cluster is consistent with the scenario described above, in which bacteria might be colonizing senescent *O. triangulata* cells associated with this cryptic group. This is supported, in addition to the lower abundance of total RNA reads associated to this cluster, by its proportionally high cell-associated bacterial load, observed as the significantly high proportion of prokaryotic rRNA in the cryptic compared to the other two groups. As the cells in the cryptic cluster do not seem to be actively feeding according to the down-regulation of key lysosome-related enzymes, colonization by bacteria would therefore seem a plausible explanation of the higher prokaryotic rRNA proportion. Finally, the consistent bacterial community composition associated to these cells could be explained by very specific community members taking part in the invasion.

Overall, our results show that scRNA-seq is technically apt to deal with small microeukaryotes, even those lacking a reference genome. Lack of a good quality annotated genome is a hindrance to infer biological state of a given population, which is the case for most aquatic microeukaryotes (Keeling et al. 2014). Therefore, while potential for detailed mechanistic understanding of functional states that are poorly represented in public databases is very limited, potential for discovery of unexpected and diverse expression consortia remains a key asset of scRNA-seq. Taken together, the combination of 18S rRNA annotation with expression-driven cell clustering holds promise as a

powerful tool to characterize community structure from natural samples beyond metabarcoding.

Acknowledgements

This work was partly financed by the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ITN project SINGEK (grant agreement no. H2020-MSCA-ITN-2015-675752), the Swedish Research Council (grant 2017-04422), Formas (grant 2019-02366) and the Academy of Finland (grant 336475). The authors would like to thank Ian Probert (Roscoff Culture Collection) for providing the strain used in this study. The authors would like to acknowledge support from the Genomics infrastructure services at Science for Life Laboratory. Single Cell Sorting and library preparation was performed at the Microbial Single Cell Genomics Facility (MSCG). Sequencing was performed by the SNP&SEQ Technology Platform. The facilities are part of the Uppsala node at the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. The Eukaryotic Single Cell Genomics unit at the Stockholm node, Science for Life Laboratory, provided instrumental advice when adapting the SmartSeq2 protocol to low volume reactions. The authors also wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Andersen, R.A. (1982) 'A light and electron microscopical investigation of *Ochromonas sphaerocystis* Matvienko (Chrysophyceae): the statospore, vegetative cell and its peripheral vesicles', *Phycologia*, 21(3), pp. 390–398.

Berges, J.A., Franklin, D.J. and Harrison, P.J. (2001) 'Evolution of an artificial seawater medium: Improvements in enriched seawater, artificial water over the last two decades', *Journal of Phycology*, 37(6), pp. 1138–1145.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, pp. 1–9.

Del Campo, J. *et al.* (2014) 'The others: our biased perspective of eukaryotic genomes', *Trends in Ecology and Evolution*, 29(5), pp. 252–259.

Cooper, E.D. *et al.* (2014) 'Metatranscriptome profiling of a harmful algal bloom', *Harmful Algae*, 37, pp. 75–83.

Dainat, J., Hereñú, D. and Pucholt, P. (2021) 'AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format'.

Deng, Z.L., Münch, P.C., Mreches, R. and McHardy, A.C. (2022) 'Rapid and accurate identification of ribosomal RNA sequences via deep learning', *Nucleic Acids Research*, 50(10), p. E60.

Edgar, R. (2016) 'SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences', *bioRxiv* [Preprint], (074161).

Forster, D. *et al.* (2016) 'Benthic protists: the under-charted majority Dominik', *FEMS Microbiology Ecology*, 92, pp. 1–11.
Geisen, S. *et al.* (2015) 'Metatranscriptomic census of active protists in soils', *ISME Journal*, 9(10), pp. 2178–2190.
Geisen, S. *et al.* (2016) 'The soil food web revisited: Diverse and widespread mycophagous soil protists', *Soil Biology and Biochemistry*, 94, pp. 10–18.
Goffeau, A. *et al.* (1996) 'Life with 6000 Genes', *Science*, 274, pp. 546–567.
Grabherr, M.G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652.
Guillou, L. *et al.* (2013) 'The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy', *Nucleic Acids Research*, 41(D1), pp. 597–604.
Haas, B.J. (2018) 'TransDecoder v5.5.0'. Available at: https://github.com/TransDecoder/TransDecoder.
Hibberd, D.J. (1977) 'Ultrastructure of cyst formation in *Ochromonas tuberculata* (Chrysophyceae)', *Journal of Phycology*, 13, pp. 309–320.
Howick, V.M. *et al.* (2019) 'The malaria cell atlas: Single parasite transcriptomes across the complete *Plasmodium* life cycle', *Science*, 365(6455).
Hsieh, T.C., Ma, K.H. and Chao, A. (2016) 'iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers)', *Methods in Ecology and Evolution*, 7(12), pp. 1451–1456.
Iniesto, M. *et al.* (2022) 'Planktonic microbial communities from microbialite-bearing lakes sampled along a salinity-alkalinity gradient', *Limnology and Oceanography*, 67(12), pp. 2718–2733.
Jamy, M. *et al.* (2022) 'Global patterns and rates of habitat transitions across the eukaryotic tree of life', *Nature Ecology and Evolution*, 6(October), pp. 1458–1470.
Jones, H.L.J. (1997) 'A classification of mixotrophic protists based on their behaviour', *Freshwater Biology*, 37(1), pp. 35–43.
Kapteyn, J., He, R., McDowell, E.T. and Gang, D.R. (2010) 'Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples', *BMC Genomics*, 11(1).
Keeling, P.J. *et al.* (2014) 'The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing', *PLoS Biology*, 12(6).
Keller, M.D., Selvin, R.C., Claus, W. and Guillard, R.R.L. (1987) 'Media for the culture of oceanic ultraphytoplankton', *Journal of Phycology*, 23(4), pp. 633–638.
Kristiansen, J. and Škaloud, P. (2017) 'Chrysophyta', in J.M. Archibald, A.G.B. Simpson, and C.H. Slamovits (eds) *Handbook of the Protists*. Springer.
Ku, C. *et al.* (2020) 'A single-cell view on alga-virus interactions reveals sequential transcriptional programs and infection states', *Science Advances*, 6(21).
Ku, C. and Sebé-Pedrós, A. (2019) 'Using single-cell transcriptomics to understand functional states and interactions in microbial eukaryotes', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786).
Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome biology*, 10(3), p. R25.
Lepere, C. *et al.* (2011) 'Whole-genome amplification (WGA) of marine photosynthetic eukaryote populations', *FEMS Microbiology Ecology*, 76(3), pp. 513–523.

Li, B. and Dewey, C.N. (2011) 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics*, 12(323), pp. 1–16.

Lie, A.A.Y. *et al.* (2018) 'A tale of two mixotrophic chrysophytes: Insights into the metabolisms of two *Ochromonas* species (Chrysophyceae) through a comparison of gene expression', *PLoS ONE*, 13(2), pp. 1–20.

Liu, Z. *et al.* (2017) 'Single-cell transcriptomics of small microbial eukaryotes: Limitations and potential', *ISME Journal*, 11(5), pp. 1282–1285.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21.

Luo, W. and Brouwer, C. (2013) 'Pathview: an R/Bioconductor package for pathway-based data integration and visualization', *Bioinformatics*, 29(14), pp. 1830–1831.

Mahata, B. *et al.* (2014) 'Single-cell RNA sequencing reveals T helper cells synthesizing steroids De Novo to contribute to immune homeostasis', *Cell Reports*, 7(4), pp. 1130–1142.

Mahé, F. *et al.* (2017) 'Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests', *Nature Ecology and Evolution*, 1(4).

Mangot, J.-F. *et al.* (2017) 'Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells', *Scientific Reports*, 7, p. 41498.

Marchetti, A. *et al.* (2012) 'Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability', *Proceedings of the National Academy of Sciences of the United States of America*, 109(6).

Marshall, W. and Laybourn-Parry, J. (2002) 'The balance between photosynthesis and grazing in Antarctic mixotrophic cryptophytes during summer', *Freshwater Biology*, 47(11), pp. 2060–2070.

Massana, R. *et al.* (2014) 'Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles.', *The ISME Journal*, 8(4), pp. 854–66.

McConnell, M.J. *et al.* (2013) 'Mosaic Copy Number Variation in Human Neurons', *Science*, 342(6158), pp. 632–638.

Mitra, A. *et al.* (2016) 'Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition: Incorporation of Diverse Mixotrophic Strategies', *Protist*, 167(2), pp. 106–120.

Moreira, D. and López-García, P. (2019) 'Time series are critical to understand microbial plankton diversity and ecology', *Molecular Ecology*, 28(5), pp. 920–922.

Nadal-Ribelles, M. *et al.* (2019) 'Sensitive high-throughput single-cell RNA-seq reveals within-clonal transcript correlations in yeast populations', *Nature Microbiology*, 4(April), pp. 683–692. Available at: http://dx.doi.org/10.1038/s41564-018-0346-9.

Navin, N. *et al.* (2011) 'Tumour evolution inferred by single-cell sequencing', *Nature*, 472(7341), pp. 90–94.

Pernice, M.C. *et al.* (2016) 'Large variability of bathypelagic microbial eukaryotic communities across the world's oceans', *ISME Journal*, 10(4), pp. 945–958.

Picelli, S. *et al.* (2014) 'Full-length RNA-seq from single cells using Smart-seq2', *Nature Protocols*, 9(1), pp. 171–181.

Poran, A. *et al.* (2017) 'Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites', *Nature*, 551(7678), pp. 95–99.

Reid, A.J. *et al.* (2018) 'Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites', *eLife*, 7, pp. 1–29.
Rinke, C. *et al.* (2013) 'Insights into the phylogeny and coding potential of microbial dark matter', *Nature*, 499(7459), pp. 431–437.
Rognes, T. *et al.* (2016) 'VSEARCH: A versatile open source tool for metagenomics', *PeerJ*, 2016(10), pp. 1–22.
Rubino, F. *et al.* (2014) 'MGkit: Metagenomic Framework For The Study Of Microbial Communities'. Available at: https://bitbucket.org/setsuna80/mgkit.
Saint, M. *et al.* (2019) 'Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation', *Nature Microbiology*, 4(3), pp. 480–491.
Simpson, A.G.B., Slamovits, C.H. and Archibald, J.M. (2017) 'Protist Diversity and Eukaryote Phylogeny', in J.M. Archibald, A.G.B. Simpson, and C.H. Slamovits (eds) *Handbook of the Protists*. 2nd edn. Cham, Switzerland: Springer, pp. 1–21.
Smith, D.R. and Keeling, P.J. (2016) 'Protists and the Wild, Wild West of Gene Expression: New Frontiers, Lawlessness, and Misfits', *Annual Review of Microbiology*, 70(1), p. annurev-micro-102215-095448.
Suter, D.M. *et al.* (2011) 'Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics', *Science*, 332(6028), pp. 472–474.
Tesson, B., Lerch, S.J.L. and Hildebrand, M. (2017) 'Characterization of a new protein family associated with the silica deposition vesicle membrane enables genetic manipulation of diatom silica', *Scientific Reports*, 7(1), pp. 1–13.
Turchinovich, A. *et al.* (2014) 'Capture and Amplification by Tailing and Switching (CATS), Anultrasensitive ligation-independent method for generation of DNA libraries for deep sequencing from picogram amounts of DNA and RNA', *RNA Biology*, 11(7), pp. 817–828.
Vass, M., Székely, A.J., Lindström, E.S. and Langenheder, S. (2020) 'Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions', *Scientific Reports*, 10(1), pp. 1–13.
Venter, J.C. *et al.* (2004) 'Environmental Genome Shotgun Sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–75.
Walker, G., Dorrell, R.G., Schlacht, A. and Dacks, J.B. (2011) 'Eukaryotic systematics: a user's guide for cell biologists and parasitologists', *Parasitology*, 138(13), pp. 1638–1663.
Wood, D.E. and Salzberg, S.L. (2014) 'Kraken: Ultrafast metagenomic sequence classification using exact alignments', *Genome Biology*, 15(3).
Worden, A.Z. *et al.* (2015) 'Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes', *Science*, 347(6223).
Xue, Y. *et al.* (2020) 'A single-parasite transcriptional atlas of *Toxoplasma gondii* reveals novel control of antigen expression', *eLife*, 9, pp. 1–27.
Yilmaz, P. *et al.* (2014) 'The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks', *Nucleic Acids Research*, 42(D), pp. 643–648.
Yoon, H.S. *et al.* (2011) 'Single-cell genomics reveals organismal interactions in uncultivated marine protists', *Science (New York, N.Y.)*, 332(6030), pp. 714–717.